ECE656-Machine Learning and Adaptive Systems
Lectures 21 & 22

M.R. Azimi, Professor

Department of Electrical and Computer Engineering
Colorado State University

Fall 2015
Fisher Linear Discriminant (1936)-Two-Class Story

Why FDA versus PCA?

- PCA finds components that are useful for representing data in lower dimensional subspace by maximizing the variance. Thus, low variance (energy) components are discarded.
- PCA is an unsupervised method i.e. data samples are unlabeled and hence no attempt is made to extract discriminatory components. Thus, the directions discarded by PCA may happen to be the most discriminatory features (e.g. discriminating O and Q).
- In contrast, FDA seeks to reduce dimensionality by preserving most discriminatory features.

Let \( \{x_p, d_p\}_{p=1}^P \), with \( x_p \in \mathbb{R}^N \) and \( d_p \in \mathbb{R} \), be a set of labeled data samples, \( P_1 \) of class \( C_1 \) (\( d_p = -1 \)) and \( P_2 \) of class \( C_2 \) (\( d_p = 1 \)) and \( P_1 + P_2 = P \). We seek to obtain a mapping,

\[ y_p = w^t x_p \]

that linearly combines samples, \( x_p \), to scalars, \( y_p \) such that class separation among these scalars is maximized. Clearly, if \( ||w|| = 1 \) this corresponds to projecting each sample onto a line and in which case we are seeking a line direction that maximizes the separation (see figure).
To find the best $w$, we use the sample mean as a measure of separation between the projected points. If we define sample mean vector of each class $C_i$, 

$$\mu_i = \frac{1}{P_i} \sum_{p=1}^{P_i} x_p, \quad i = 1, 2$$

then the sample mean for the projected points is,

$$m_i = \frac{1}{P_i} \sum_{p=1}^{P_i} y_p = \frac{1}{P_i} \sum_{p=1}^{P_i} w^t x_p = w^t \mu_i, \quad i = 1, 2$$

Note that the training samples are arranged in two different subsets for classes of $C_1$ and $C_2$. We use the distance between the projected means as our measure, i.e.

$$|m_1 - m_2| = |w^t (\mu_1 - \mu_2)|$$

However, the distance between projected means alone is not a good measure since it does not account for the standard deviation within classes.
Fisher suggested maximizing the difference between the means, normalized by a measure of the within-class scatter of the projected samples given by, \((\sigma_1^2 + \sigma_2^2)\), where \(\sigma_i^2\) is the sample variance of \(y\) sample in class \(C_i\). Thus, the Fisher measure to maximize wrt \(w\) is,

\[
J(w) = \frac{|m_1 - m_2|^2}{\sigma_1^2 + \sigma_2^2}
\]

To find the optimum \(w\) we must express \(J(w)\) as a function of \(w\). We define scatter matrices \(S_i\) as

\[
S_i = \sum_{p=1}^{P_i} (x_p - \mu_i)(x_p - \mu_i)^t
\]

i.e. within class sample covariance matrices (a measure of the scatter in feature space \(x\) for class \(C_i\)). Also, define the within-class scatter matrix

\[
S_W = S_1 + S_2
\]
Fisher Discriminant Analysis (FDA)  

The variance (or scatter) of the projected samples in $C_i$ can be expressed as,

$$\sigma_i^2 = \sum_{p=1}^{P_i} (y_p - m_i)^2 = \sum_{p=1}^{P_i} (w^t x_p - w^t \mu_i)^2$$

$$= \sum_{p=1}^{P_i} w^t (x_p - \mu_i)(x_p - \mu_i)^t w = w^t S_i w$$

Thus, the denominator of $J(w)$, or the total within-class scatter, becomes

$$(\sigma_1^2 + \sigma_2^2) = w^t (S_1 + S_2) w = w^t S_W w.$$ 

Also, we can express the numerator of $J(w)$ as

$$(m_1 - m_2)^2 = (w^t \mu_1 - w^t \mu_2)^2 = w^t (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t w = w^t S_B w$$

where matrix $S_B$ is called the between-class scatter matrix. Note that though $S_W$ is symmetric, PD and nonsingular (when $P_i \gg N$), matrix $S_B$ is symmetric, PD but singular (rank one).

Now, the Fisher criterion in terms of $w$ becomes,

$$J(w) = \frac{w^t S_B w}{w^t S_W w}$$

i.e. generalized Rayleigh quotient.

An important property to notice about the Fisher measure $J(w)$ is that it is invariant to scaling of the weight vector i.e. $w \rightarrow aw$. Hence, we choose a $w$ such that $w^t S_W w = 1$. 

M.R. Azimi  Machine Learning and Adaptive Systems
Fisher Linear Discriminant-Cont.

This leads to the following constrained optimization

Maximize \( w^t S_B w \) subject to \( w^t S_W w = 1 \).

which can be represented by Lagrangian cost function,

\[
\max_w \left\{ w^t S_B w - \lambda (w^t S_W w - 1) \right\}
\]

The solution of this maximization gives the following *generalized eigenvalue problem*,

\[ S_B w = \lambda S_W w \]

Alternatively,

\[ S_W^{-1} S_B w = \lambda w \]

**Remarks:**

- Note that it is not necessary to solve for eigenvalues and eigenvectors of \( S_W^{-1} S_B \) since \( S_B w \) is always in the direction of \( (\mu_1 - \mu_2) \). Thus, due to scale invariance, we can write the solution for optimum \( w \) as \( w^* = S_W^{-1} (\mu_1 - \mu_2) \).
- Using Fisher linear discriminant classification is reduced from \( N \)-D problem to a 1-D problem.
- This mapping is many-to-one and cannot reduce minimum achievable error rate.
Example:

Compute the LDA projection for the following 2-D data subsets

\[ X_1 = \{(4, 1), (2, 4), (2, 3), (3, 6), (4, 4)\} \] in class \( C_1 \)
\[ X_2 = \{(9, 10), (6, 8), (9, 5), (8, 7), (10, 8)\} \] in class \( C_2 \).

The class scatter matrices are:

\[
S_1 = \begin{bmatrix}
0.8 & -0.4 \\
-0.4 & 2.64
\end{bmatrix}
\quad \text{and} \quad
S_2 = \begin{bmatrix}
1.84 & -0.04 \\
-0.04 & 2.64
\end{bmatrix}
\]

Also, \( \mu_1 = [3 \ 3.6]^t \) and \( \mu_2 = [8.4 \ 7.6]^t \).

Thus, we get

\[
S_B = \begin{bmatrix}
29.16 & 21.6 \\
21.6 & 16
\end{bmatrix}
\quad \text{and} \quad
S_W = \begin{bmatrix}
2.64 & -0.44 \\
-0.44 & 5.28
\end{bmatrix}.
\]

First we use the direct approach by solving eigenvalues and eigenvectors of

\[
S_W^{-1} S_B w = \lambda w. \quad \text{The eigenvector that gives the largest } J(w) \text{ is } w^* = [0.91 \ 0.39]^t.
\]

Alternatively, we use the result in Remark 1 i.e.

\[
w^* = S_W^{-1} (\mu_1 - \mu_2) = [-0.91 \ -0.39]^t.
\]
Multiple Discriminant Analysis

For $M$-class problem, instead of one projection $y$, we will now seek $(M - 1)$ projections, $y^{(1)}, y^{(2)}, \ldots, y^{(M-1)}$ by means of projection vectors $w_i, i \in [1, M - 1]$ (or DFs) arranged by columns into a projection matrix $W = [w_1 w_2 \cdots w_{M-1}]$. That is,

$$y_p^{(i)} = w_i^t x_p \quad \text{or} \quad y_p = W^t x_p.$$  

Following the same procedure as in the LDA, we have

$$S_i = \sum_{p=1}^{P_i} (x_p - \mu_i)(x_p - \mu_i)^t$$

where $\mu_i = \frac{1}{P_i} \sum_{p=1}^{P_i} x_p$. The within-class scatter matrix becomes,

$$S_W = \sum_{i=1}^{M} \sum_{p=1}^{P_i} (x_p - \mu_i)(x_p - \mu_i)^t.$$

The between-class scatter matrix in this case is

$$S_B = \sum_{i=1}^{M} P_i (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^t$$

where $\bar{\mu} = \frac{1}{P} \sum_{p=1}^{P} x_p = \frac{1}{P} \sum_{i=1}^{M} P_i \mu_i$.

Similarly, we define the mean vector and scatter matrices for the projected samples as,

$$\tilde{S}_W = \sum_{i=1}^{M} \sum_{p=1}^{P_i} (y_p - \tilde{\mu}_i)(y_p - \tilde{\mu}_i)^t,$$

where $\tilde{\mu}_i = \frac{1}{P_i} \sum_{p=1}^{P_i} y_p$. 

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure}
\caption{Illustration of Multiple Discriminant Analysis.}
\end{figure}
Multiple Discriminant Analysis-Cont.

Also,
$$\tilde{S}_B = \sum_{i=1}^{M} P_i (\tilde{\mu}_i - \tilde{\mu}) (\tilde{\mu}_i - \tilde{\mu})^t.$$  

where
$$\tilde{\mu} = \frac{1}{P} \sum_{p=1}^{P} y_p = \frac{1}{P} \sum_{i=1}^{M} P_i \tilde{\mu}_i.$$  

It is straightforward to show that,
$$\tilde{S}_W = W^t S_W W \quad \text{and} \quad \tilde{S}_B = W^t S_B W.$$  

Since the projection is no longer a scalar (it has $M - 1$ dimensions), we use the determinant of the scatter matrices to obtain a scalar objective function to find optimum $W$. This is,
$$J(W) = \frac{|W^t S_B W|}{|W^t S_W W|}.$$  

It can be shown that the optimal projection matrix $W^*$ is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem,
$$W^* = [w_1^* w_2^* \cdots w_{M-1}^*] = \text{argmax} \frac{|W^t S_B W|}{|W^t S_W W|} \implies (S_B - \lambda_i S_W) w_i^* = 0.$$  

Remarks:

- Matrix $S_B$ is the sum of $M$ rank-1 matrices and the mean vectors are constrained by $\tilde{\mu} = \frac{1}{P} \sum_{i=1}^{M} P_i \tilde{\mu}_i$. Thus, matrix $S_B$ will be of rank $M - 1$ or less. This means that only $M - 1$ of the eigenvalues $\lambda_i$ will be non-zero.
The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $S^{-1}_W S_B$.

EVD can be avoided by computing root of $|S_B - \lambda_i S_W| = 0$.

LDA can be derived as the Maximum Likelihood method for the case of normal class-conditional densities with equal covariance matrices.

LDA produces at most $M - 1$ feature projections (principal components of $S_B$). If the classification error estimates establish that more features are needed, some other method must be employed to provide those additional features.

LDA is a parametric method (it assumes unimodal Gaussian likelihoods). If distributions are non-Gaussian, the LDA projections may not preserve complex structures in the data needed for classification.
Multidimensional Scaling (MDS)

**Goal:**

Find a low-dimensional representation set of points which are most closely consistent (under some cost function) with a measured set dissimilarities $D \in \mathbb{R}^{N \times N}$ (or pre-distance matrix). That is, construct a set of low dimensional points whose interpoint Euclidean distances match those in $D$ closely. Note that here the original data points are not directly observable, only their distances (e.g., construct geographical map of CO based upon road distances of towns).

There are many cost functions that can be used leading to different MDS algorithms. Here, we focus on classical MDS.

Let $\hat{D}$ denote the distance matrix between the points in the low-dimensional space and $\rho(D, \hat{D})$ be an appropriate cost function. One very convenient cost function is called the “STRAIN criterion” given by

$$\rho(D, \hat{D}) = \|P_c(D - \hat{D})P_c\|_F^2,$$

where $\|A\|_F^2 = tr(AA^H)$ is the Frobenius norm of matrix $A$, and $P_c = I - \frac{1}{N} 1 1^t$ represents a centering matrix. Note that $P_cX$ subtracts the row mean from each row of matrix $X$ while $XP_c$ subtracts the column mean from each column of matrix $X$.

Using this cost function one would like to match variations in distance, rather than the values themselves; this makes it invariant to translation, rotation, and reflection. The solution of the above problem is given in the next theorem.
Definition 1: A distance matrix \( D \in \mathcal{R}^{L \times L} \) has the properties (a) \( d_{ii} = 0 \), (b) \( d_{ij} \geq 0 \), \( i \neq j \), and (c) \( d_{ij} = d_{ji} \).

Definition 2: A distance matrix \( D \) is called Euclidean if there exists a configuration of points \( x_1, \ldots, x_L \in \mathbb{R}^N \) whose interpoint distances are given by \( D \) i.e. 

\[
d_{ij} = (x_i - x_j)^t (x_i - x_j) = \|x_i - x_j\|^2.
\]

Theorem:

Let \( D \) be a distance matrix and define \( \Delta = -\frac{1}{2} P_c D P_c \). Then, \( D \) is Euclidean iff \( \Delta \) is PSD. In particular, the following hold:

- **Matrix** \( \Delta \) is the Gram matrix for a mean-centered configuration with interpoint distances given by \( D \).
- If \( D \) is the Euclidean distance matrix for configuration \( X \) then 

  \[
  \delta_{ij} = (x_i - \mu_x)^t (x_j - \mu_x) \quad \text{where} \quad \mu_x = \frac{1}{L} \sum_{i=1}^{L} x_i. 
  \]

  In matrix form 

  \[
  \Delta = P_c X (P_c X)^t, \quad \text{so} \quad \Delta \geq 0 \text{ i.e. the centered inner product matrix for } X.
  \]
- Conversely, if \( \Delta \) is PSD then a configuration corresponding to \( \Delta \) can be constructed as follows. Let \( \Delta = U \Lambda U^t \) be the spectral decomposition of \( \Delta \) with 

  \[
  \Lambda = Diag[\lambda_1 \cdots \lambda_L].
  \]

  Now, since \( \Delta \) is a Gram matrix, \( \Delta = YY^t \), where 

  \[
  Y = U \Lambda^{\frac{1}{2}}. 
  \]

  Note that the points in configuration \( Y \) have also interpoint distance \( D \) but with mean of zero. Additionally, they are related to the eigenvectors of \( \Delta \).
- **Rows of** \( Y \) **correspond to the new points with same interpoint distances given by** \( D \).
To get a low $M$-dimensional embedding corresponding to $M$ non-negative eigenvalues $\lambda_i$s, we can discard the last $N - M$ columns of $Y$.

To see the relation with PCA, we use PCA to project $Y$ onto its $M$ principal coordinates i.e. $Y \Psi$, where $\Psi \in \mathcal{R}^{N \times M}$ and the columns of $\Psi = [\psi_1, \cdots, \psi_M]$ are the first $M$ eigenvectors of $Y^tY$ (covariance matrix). That is,

$$Y^tY \psi_i = \gamma_i \psi_i$$

Now, using $Y = U \Lambda^{1/2}$ yields,

$$(U \Lambda^{1/2})^t U \Lambda^{1/2} \psi_i = \gamma_i \psi_i$$

$$\Lambda^{1/2} U^t U \Lambda^{1/2} \psi_i = \gamma_i \psi_i \quad \text{or} \quad \Lambda \psi_i = \gamma_i \psi_i \quad \text{since} \quad U^t U = I.$$

which suggests that $\psi_i = e_i$ and $\gamma_i = \lambda_i$ where $e_i$ is the $i^{th}$ standard basis vector for $\mathcal{R}^L$. Then, $Y \Psi = Y[e_1 \cdots e_M]$ which corresponds to the first $M$ columns of $Y$.

Remarks:

1. The truncation step is indeed equivalent to projecting $Y$ onto its first $M$ principal components. Moreover, $\lambda_i$s reveal relative importance of each dimension as they are the same as those found by PCA.

2. MDS produced a linear embedding that preserved all pairwise distances. However, it cannot be used to nonlinear manifold learning as the Euclidean distance measurements do not respect topology of the manifold.
Fisher Discriminant Analysis (FDA)  Multiple Discriminant Analysis (MDA)  Multidimensional Scaling (MDS)

MDS Example

Consider a 7x7 distance matrix

\[
D = \begin{bmatrix}
0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & 1 \\
0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & 0 \\
0 & 1 & \sqrt{3} & 2 & 1 & 0 & 0 \\
0 & 1 & \sqrt{3} & 1 & 1 & 0 & 0 \\
0 & 1 & \sqrt{3} & 2 & 1 & 0 & 0 \\
0 & 1 & \sqrt{3} & 1 & 1 & 0 & 0 \\
0 & 1 & \sqrt{3} & 2 & 1 & 0 & 0 \\
\end{bmatrix}
\]

Constructing the \( \Delta \) matrix gives

\[
\Delta = -\frac{1}{2} P_c D P_c = \begin{bmatrix}
2 & 1 & -1 & -2 & -1 & 1 & 0 \\
2 & 1 & -1 & -2 & -1 & 0 & 0 \\
2 & 1 & -1 & -2 & 0 & 0 & 0 \\
2 & 1 & -1 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

Columns of \( \Delta \) are linearly dependent. It can be shown that \( \delta_3 = \delta_2 - \delta_1 \), \( \delta_4 = -\delta_1 \), \( \delta_5 = -\delta_2 \), \( \delta_6 = \delta_1 - \delta_2 \), and \( \delta_7 = 0 \). Thus, the rank of \( \Delta \) is 2. Hence a configuration exactly fitting the distance matrix \( D \) can be constructed in \( M = 2 \) dimensions. Eigenvalues of \( \Delta \) are \( \lambda_1 = \lambda_2 = 3 \) and \( \lambda_3 \cdots = \lambda_7 = 0 \). Using the associated eigenvectors,

\[
u_1 = [\frac{\sqrt{3}}{2}, \frac{\sqrt{3}}{2}, 0, -\frac{\sqrt{3}}{2}, -\frac{\sqrt{3}}{2}, 0, 0]^t \quad \text{and} \quad \nu_2 = [\frac{1}{2}, -\frac{1}{2}, -1, -\frac{1}{2}, \frac{1}{2}, 1, 0]^t.
\]

The coordinates of 7 points are \( [\frac{\sqrt{3}}{2}, \frac{1}{2}]^t, [\frac{\sqrt{3}}{2}, -\frac{1}{2}]^t, [0, -1]^t, [-\frac{\sqrt{3}}{2}, -\frac{1}{2}]^t, [-\frac{\sqrt{3}}{2}, \frac{1}{2}]^t, [0, 1]^t, \text{and} \ [0, 0]^t \). Note that the mean vector of these points is \( [0, 0]^t \).